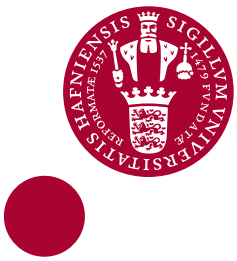

Cluster analysis of European Y-chromosomal STR haplotypes using the discrete Laplace method

Andersen, Mikkel Meyer; Eriksen, Poul Svante; Morling, Niels

Published in:
Forensic science international. Genetics

DOI:
10.1016/j.fsigen.2014.03.016

Publication date:
2014

Forensic Science International: Genetics 11 (2014) 182–194

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Cluster analysis of European Y-chromosomal STR haplotypes using the discrete Laplace method

Mikkel Meyer Andersen [a,*], Poul Svante Eriksen [a,1], Niels Morling [b,2]

[a] Department of Mathematical Sciences, Aalborg University, Denmark
[b] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 13 November 2013
Received in revised form 20 February 2014
Accepted 31 March 2014

Keywords:
Haplotype frequency
AMOVA
Population substructure
Population homogeneity
Subpopulation

ABSTRACT

The European Y-chromosomal short tandem repeat (STR) haplotype distribution has previously been analysed in various ways. Here, we introduce a new way of analysing population substructure using a new method based on clustering within the discrete Laplace exponential family that models the probability distribution of the Y-STR haplotypes. Creating a consistent statistical model of the haplotypes enables us to perform a wide range of analyses. Previously, haplotype frequency estimation using the discrete Laplace method has been validated. In this paper we investigate how the discrete Laplace method can be used for cluster analysis to further validate the discrete Laplace method. A very important practical fact is that the calculations can be performed on a normal computer.

We identified two sub-clusters of the Eastern and Western European Y-STR haplotypes similar to results of previous studies. We also compared pairwise distances (between geographically separated samples) with those obtained using the AMOVA method and found good agreement. Further analyses that are impossible with AMOVA were made using the discrete Laplace method: analysis of the homogeneity in two different ways and calculating marginal STR distributions. We found that the Y-STR haplotypes from e.g. Finland were relatively homogeneous as opposed to the relatively heterogeneous Y-STR haplotypes from e.g. Lublin, Eastern Poland and Berlin, Germany. We demonstrated that the observed distributions of alleles at each locus were similar to the expected ones.

We also compared pairwise distances between geographically separated samples from Africa with those obtained using the AMOVA method and found good agreement.

1. Introduction

Recent historical events in the European Y-chromosomal short tandem repeat (Y-STR) haplotype distribution were analysed by Roewer et al. [1] based upon a database with approximately 12,700 Y-STR profiles from 91 different locations in Europe. The analysis was performed by means of AMOVA [2], which is a cluster analysis method based upon molecular variance. In this paper, we analysed the same data using a new method based on a combination of multivariate, marginally independent, discrete Laplace distributions (called 'discrete Laplace method') as described by Andersen et al. [3,4]. We demonstrate how to use the discrete Laplace method for making inference in Y-STR haplotype databases.

The AMOVA method [2] is widely used in population and forensic genetics. The AMOVA method introduced the molecular variance measure Φ_{ST} that is an analogue to Wright's F_{ST}. Φ_{ST} is based on the detectable evolutionary distances between individual haplotypes. When a population consists of different strata (for example geographically separated sampling locations), AMOVA can be used to infer stratification through non-parametric cluster analysis of the Φ_{ST} distances.

Whereas the AMOVA method performs non-parametric cluster analysis of the Φ_{ST} distances, the discrete Laplace method by Andersen et al. [3] models the probability distribution of the Y-STR haplotypes. This makes it possible to perform much more detailed inference, e.g. estimating haplotype frequencies, model based cluster analysis, analysis of population homogeneity and comparing the observed distribution of STR alleles at each locus to the expected one.

* Corresponding author at: Fredrik Bajers Vej 7G, DK-9220 Aalborg East, Denmark. Tel.: +45 99408800.
E-mail addresses: mikl@math.aau.dk (M.M. Andersen), svante@math.aau.dk (P.S. Eriksen), niels.morling@sund.ku.dk (N. Morling).
[1] Address: Fredrik Bajers Vej 7G, DK-9220 Aalborg East, Denmark. Tel.: +45 99408800.
[2] Address: Frederik V's Vej 11, DK-2100 Copenhagen East, Denmark. Tel.: +45 35326115.

http://dx.doi.org/10.1016/j.fsigen.2014.03.016
1872-4973/

Estimating haplotype frequencies is central in forensic genetics as this is required to calculate the weight of genetic evidence as a likelihood ratio (*LR*) [3,5,6]. Hence, much attention in the forensic community is on estimating haplotype frequencies. As with any other estimation, a statistical model for doing so is recommended.

In [3], the discrete Laplace method was compared to Brenner's κ method [7] for estimating haplotype frequencies and the discrete Laplace method was found to have lower prediction error than Brenner's κ method for datasets from a range of different simulated populations.

As Brenner's κ method was used for estimating haplotype frequencies, AMOVA can be used for cluster analysis. The discrete Laplace method can be used for both as will be demonstrated for cluster analysis in this paper. As the discrete Laplace method is a consistent, statistical model, it can also be used for other analyses, e.g. population homogeneity, the distribution of STR alleles and most likely mixture separation. This shows how valuable a consistent statistical model is as specialised ad-hoc methods are not needed for each kind of analysis. Still, model validation is essential. In this paper, the cluster analysis abilities of the discrete Laplace model were investigated.

We note that the calculations can be performed on a normal computer.

2. Method

Assume that we have S different strata (for example sample locations), each with n_s individuals for $s \in \{1, 2, \ldots, S\}$, and that there are $n = \sum_{s=1}^{S} n_s$ individuals in total. Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ be the r loci Y-STR haplotype for the i'th individual for $i \in \{1, 2, \ldots, n\}$.

Assume that there are c subpopulations and that $\tau_j = P(\text{From subpopulation } j)$ is the a priori probability of a haplotype originated from the jth subpopulation for $j = 1, 2, \ldots, c$. Then

$$P(\text{Haplotype} = x | \text{From subpopulation } j) \tag{1}$$

is modelled by assuming independent discrete Laplace distributions on loci as described in Appendix A and [3]. The parameters of the model can be estimated using the `R` [8] library `disclapmix` [9] as is also shown in Appendix A.

The haplotype frequency is obtained by summing the contribution from each subpopulation, such that

$$P(\text{Haplotype} = x) = \sum_{j=1}^{c} \tau_j P(\text{Haplotype} = x | \text{From subpopulation } j). \tag{2}$$

By using Bayes theorem, we have that

$$P(\text{From subpopulation } j | \text{Haplotype} = x) = \frac{\tau_j P(\text{Haplotype} = x | \text{From subpopulation } j)}{P(\text{Haplotype} = x)}, \tag{3}$$

which can be used for cluster analysis.

2.1. Model based cluster analysis

Let $v_{ij} = P(\text{From subpopulation } j | \text{Haplotype} = x_i)$. Hence, given the haplotype of individual i, v_{ij} is the probability that the ith individual originates from the jth subpopulation. Let $\hat{v}_{ij}$ denote an estimate of v_{ij}. In this section, we analyse the $\hat{v}_{ij}$ values in a number of different ways.

To measure a distance between two subpopulations, a naïve approach of taking the minimum number of mutations between the central haplotype of the subpopulations, $\hat{y}_j$, was initially tried. Because this resulted in a large number of ties, a more sophisticated method based on the symmetrized Kullback–Leibler divergence (using the discrete Laplace method) was used. This distance measure is described in Appendix B. The distance between two subpopulations, j_1 and j_2, is denoted by

$$\mathrm{KL}(j_1, j_2). \tag{4}$$

Now, we have a distance measure between subpopulations, and we introduce a summary of the $\hat{v}_{ij}$ values for each stratum, s, and each subpopulation, j. Let I_s be the indices for the individuals in the sth stratum and let

$$w_{sj} = n_s^{-1} \sum_{i \in I_s} \hat{v}_{ij} \tag{5}$$

be the sth stratum's mean probability of originating from the jth subpopulation for $s \in \{1, 2, \ldots, S\}$ and $j \in \{1, 2, \ldots, c\}$. Note, that

$$w_{s+} = \sum_{j=1}^{c} w_{sj} = 1 \quad \text{and} \quad w_{+j} = \sum_{s=1}^{S} w_{sj} = \hat{\tau}_j. \tag{6}$$

The distance between two subpopulations can be used for constructing complete hierarchical clustering [10,11] (such that the distance between two subpopulations is the maximum distance between their individual haplotypes) of the central haplotype y_j of subpopulation for $j \in \{1, 2, \ldots, c\}$ using $\mathrm{KL}(j_1, j_2)$ given in Eq. (4) as the distance measure. This will be used for the dendograms of subpopulation distances in the following. The leaves (subpopulations) of the dendograms were reordered using the `R` [8] library `seriation` [12,13] with the `OLO` method [12]. The labels are the central haplotype of the corresponding subpopulation and the predicted haplogroup from http://www.yhrd.org release 44 [14,15].

2.1.1. Pairwise distances

A distance metric between stratum (sample location) s and t can be defined as follows. Let

$$\delta(s, t) = \sum_{j=1}^{c} (w_{sj} - w_{tj})^2 \tag{7}$$

be the pairwise (L_2) distance between stratum s and stratum t using the mean estimated subpopulation affiliations w_{sj} and w_{tj} introduced in Eq. (5). This is the squared Euclidean distance between vector $(w_{s1}, w_{s2}, \ldots, w_{sr})$ and vector $(w_{t1}, w_{t2}, \ldots, w_{tr})$. This can for example be used for hierarchical clustering.

3. Results

The primary dataset analysed is a European 7-loci Y-STR database from the year of 2004 consisting of 12,727 individuals in 91 strata (European sample locations). This dataset was first analysed by Roewer et al. [1] using AMOVA [2] among other methods. The 7 Y-STR loci were `DYS19`, `DYS389I`, `DYS389II`, `DYS390`, `DYS391`, `DYS392` and `DYS393`. The alleles at `DYS389II` were replaced by `DYS389II` minus `DYS389I` [16].

The data were best described by a discrete Laplace model with 40 subpopulations as described in more detail in Appendix A and [3].

In Fig. 1, the number of times that a haplotype was observed was compared to the estimated haplotype frequency using the discrete Laplace method. Haplotype frequency estimation using the discrete Laplace method was performed as described in Appendix A.

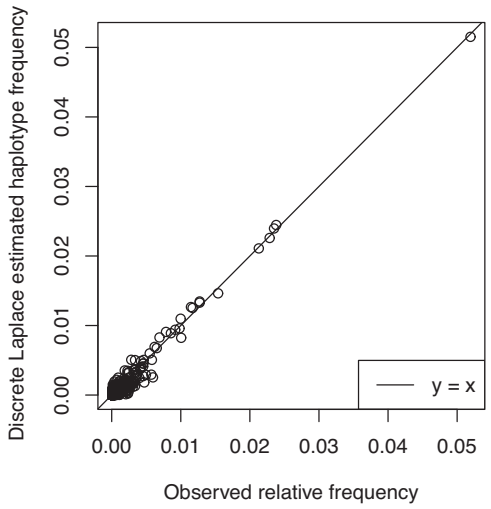

Fig. 1. Comparison of (1) the relative frequency of a haplotype (number of times it has been observed divided by the database size) and (2) the estimated haplotype frequency using the discrete Laplace method. Note, that for frequently observed haplotypes, the estimated haplotype frequency using the discrete Laplace method is close to the relative frequency.

In Fig. 2, a map of Europe with the w_{sj} values for all subpopulations $j \in \{1, 2, \ldots, c\}$ at each stratum, s (sampling locations), is shown. We refer to A for how to calculate this in R.

The majority of the central haplotypes of the subpopulations were close to each other. To better visualise the subpopulations, those with central haplotypes close to each other were assembled into mega clusters as described in Appendix A.

Motivated by Roewer et al. [1], two mega clusters were made based on the $\mathrm{KL}(j_1, j_2)$ distances between the central haplotype of subpopulations. Looking at the resulting u_{st} values on a European map as shown in Fig. 3, it seems as if an Eastern and a Western European population emerge.

If four mega clusters were chosen, a map as shown in Fig. 4 was obtained. It was now possible to identify Northern (Scandinavia), Southern (near the Balkan Peninsula), Eastern and Western European populations.

3.1. Pairwise distances

Hierarchical clustering can be done using a pairwise distance between stratum s and t as given in Eq. (7). A dendogram of these pairwise distances is shown in Fig. 5. For comparison, see Fig. 6 for hierarchical clustering of the pairwise Φ_{ST} distances calculated with Arlequin version 3.5 [18] that uses the AMOVA method by Excoffier et al. [2]. There was a very high concordance that will be analysed further below.

These pairwise distances can be compared as shown in Fig. 7. There is a strong correlation of 0.90 (Pearson's correlation coefficient, $p < 10^{-15}$) between the Φ_{ST} values and the $\delta(s, t)$ values even though they are calculated in two very different ways.

In [22], 10-locus Y-STR profiles for 2736 African individuals from 26 different strata (populations) were investigated. The 10-loci were DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438 and DYS439. The alleles at DYS389II were replaced by DYS389II minus DYS389I [16]. Removing profiles with duplications, missing or intermediate alleles, 2631 profiles remained for analysis. Calculating the pairwise $\delta(s, t)$ distances given in Eq. (7) between all population pairs and comparing those to the pairwise Φ_{ST} distances calculated with Arlequin version 3.5 [18] that uses the AMOVA method by Excoffier et al. [2], a high correlation of 0.82 (Pearson's correlation coefficient, $p < 10^{-15}$) was obtained.

3.2. Population homogeneity

In this section, we focus on two different homogeneity measures for strata and exemplify these measures by looking at three strata (sample locations). First, let

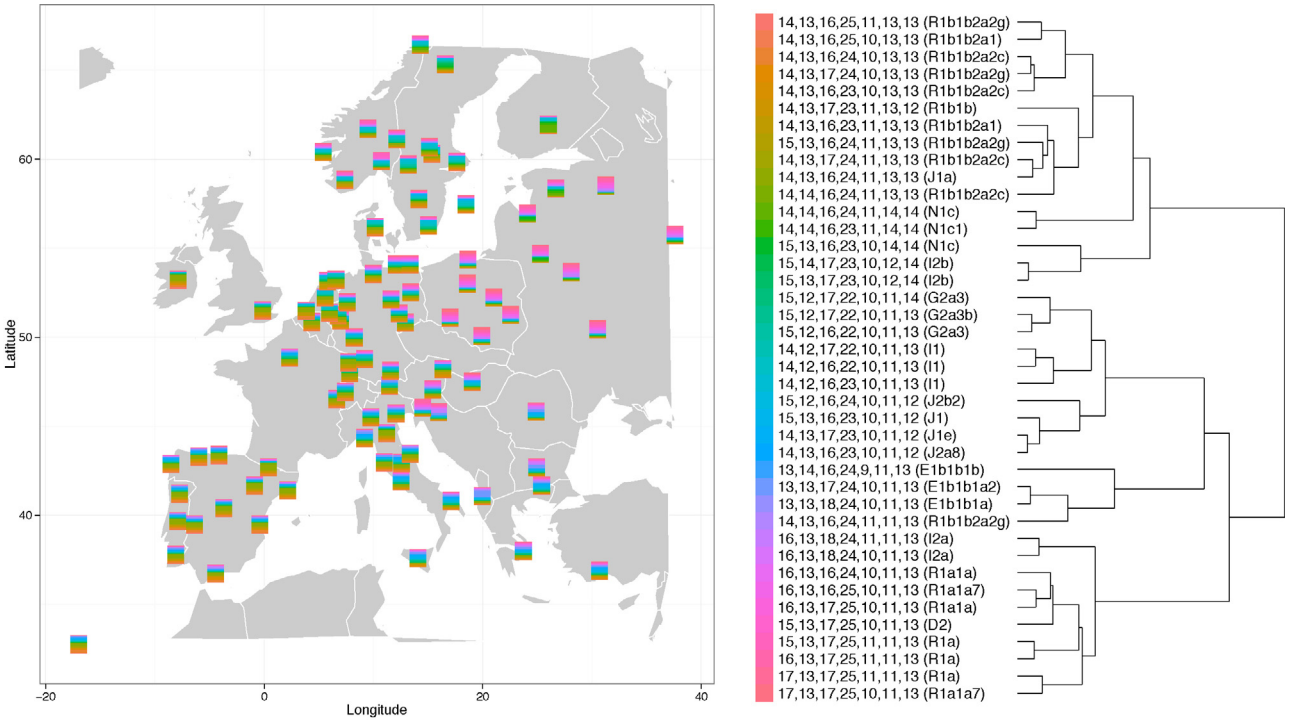

Fig. 2. Map of Europe with the w_{sj} values for all subpopulations $j \in \{1, 2, \ldots, c\}$ at each stratum s (sampling locations). A dendogram (refer to Section 2.1) is shown together with the colours for each of the $c = 40$ subpopulations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

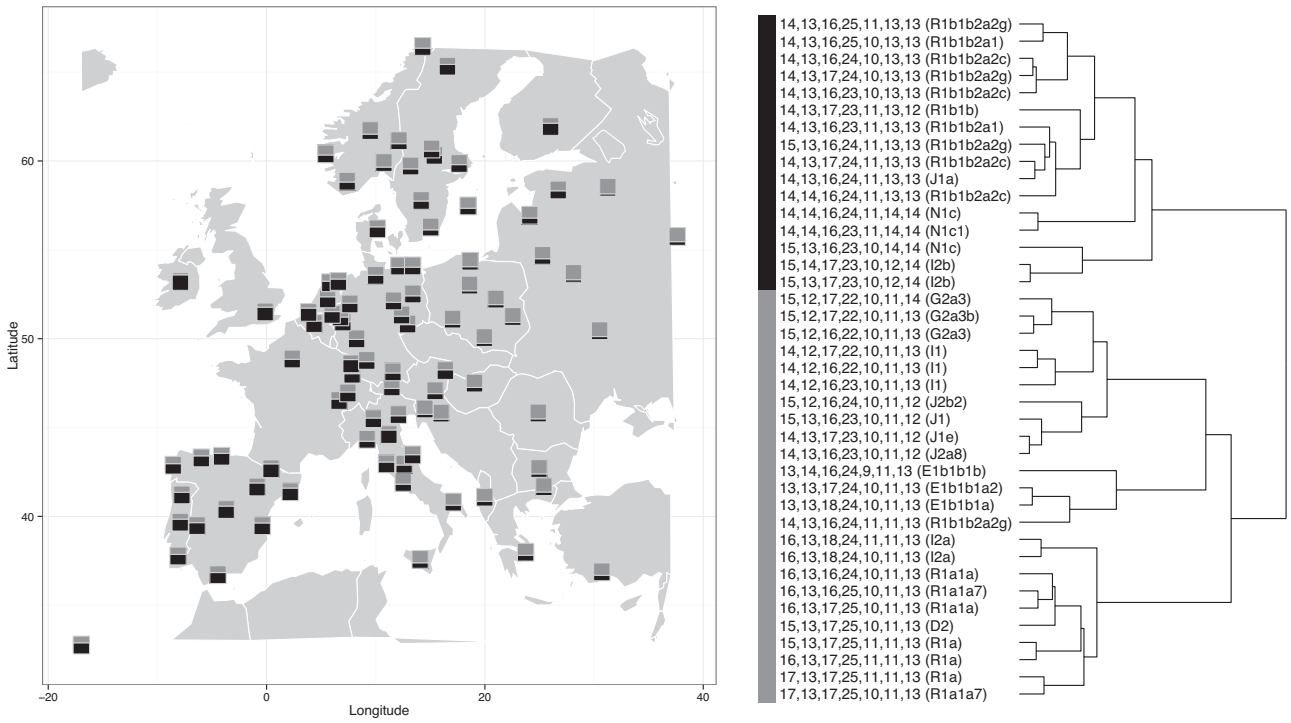

Fig. 3. Map of Europe with the u_{st} values for all mega clusters $t \in \{1, 2\}$ at each stratum s (sampling locations). Please refer to Section 2.1 for a description of the hierarchical clustering and the labels of the central haplotypes. The partitions are indicated by black and grey.

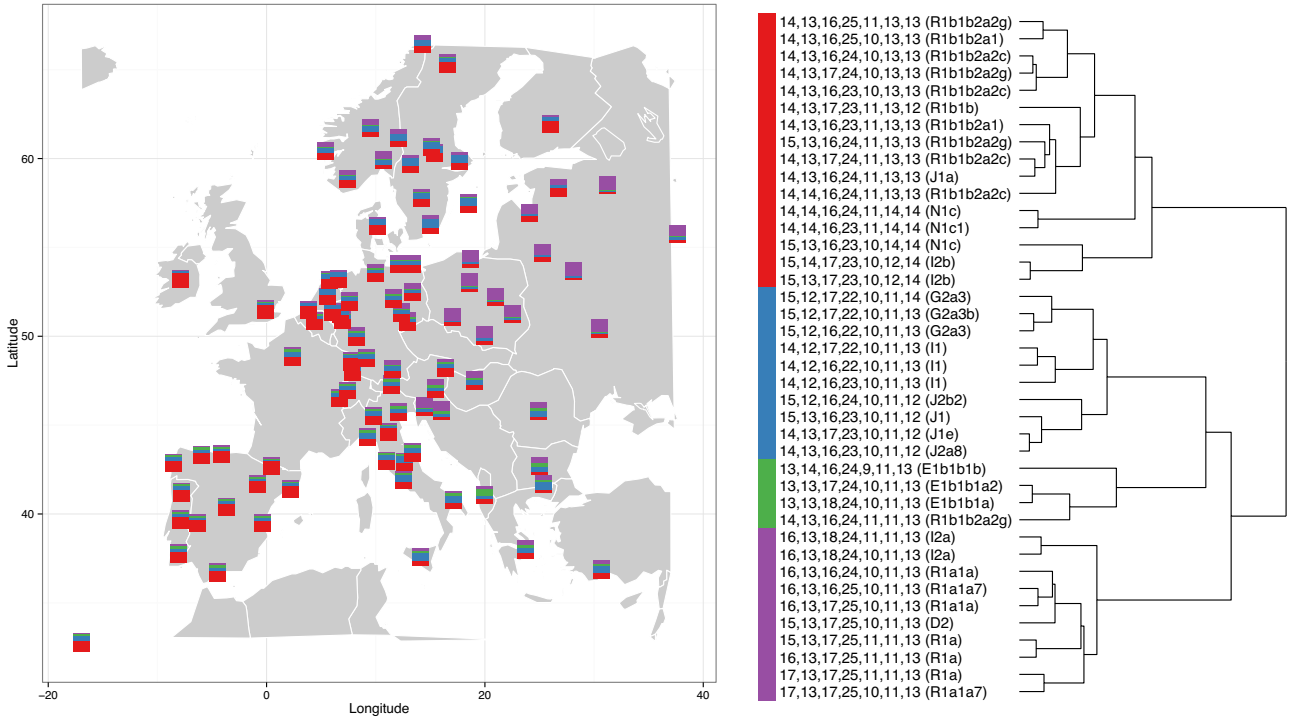

Fig. 4. Map of Europe with the u_{st} values for all mega clusters $t \in \{1, 2, 3, 4\}$ at each stratum s (sampling locations). Please refer to Section 2.1 for a description of the hierarchical clustering and the labels of the central haplotypes. The partitions are indicated by colours. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

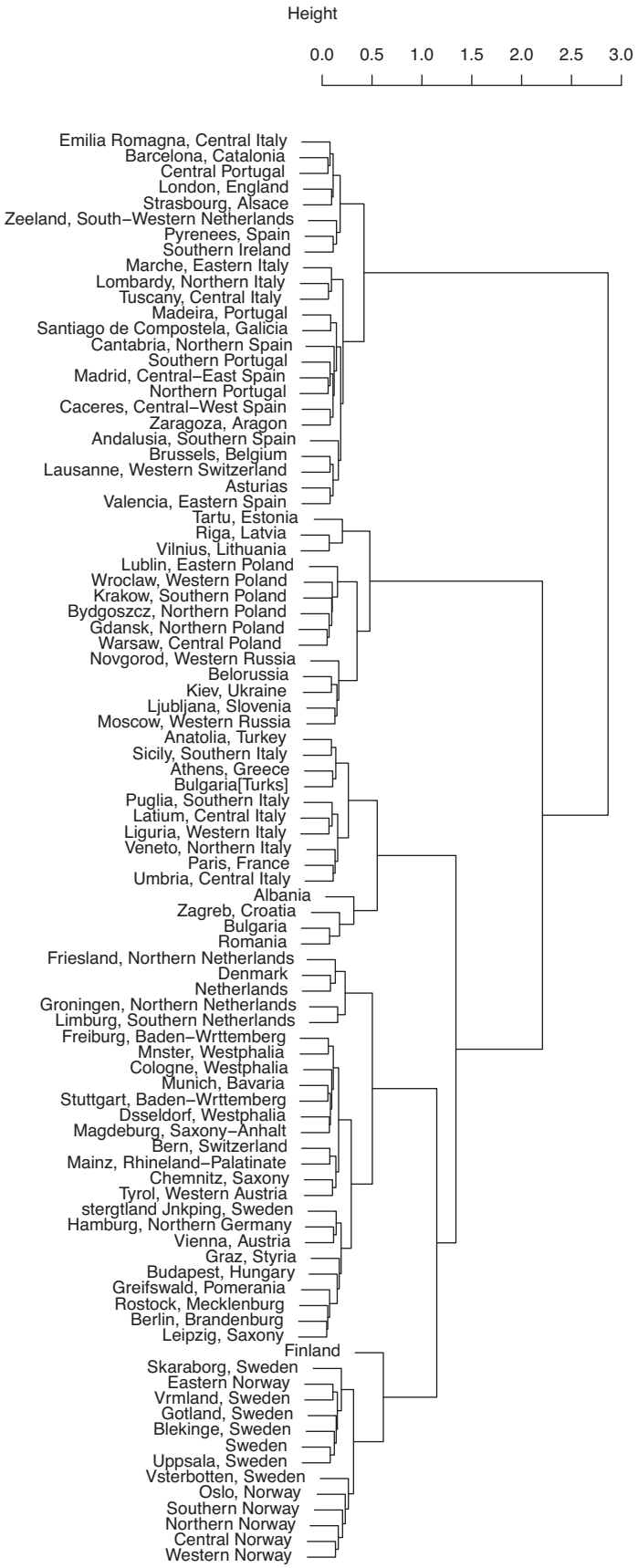

Fig. 5. A dendogram of $\delta(s,t)$ distances (using the discrete Laplace method) between all pairs of strata as defined in Eq. (7). The dendogram was made by hierarchical clustering using Ward's method [17].

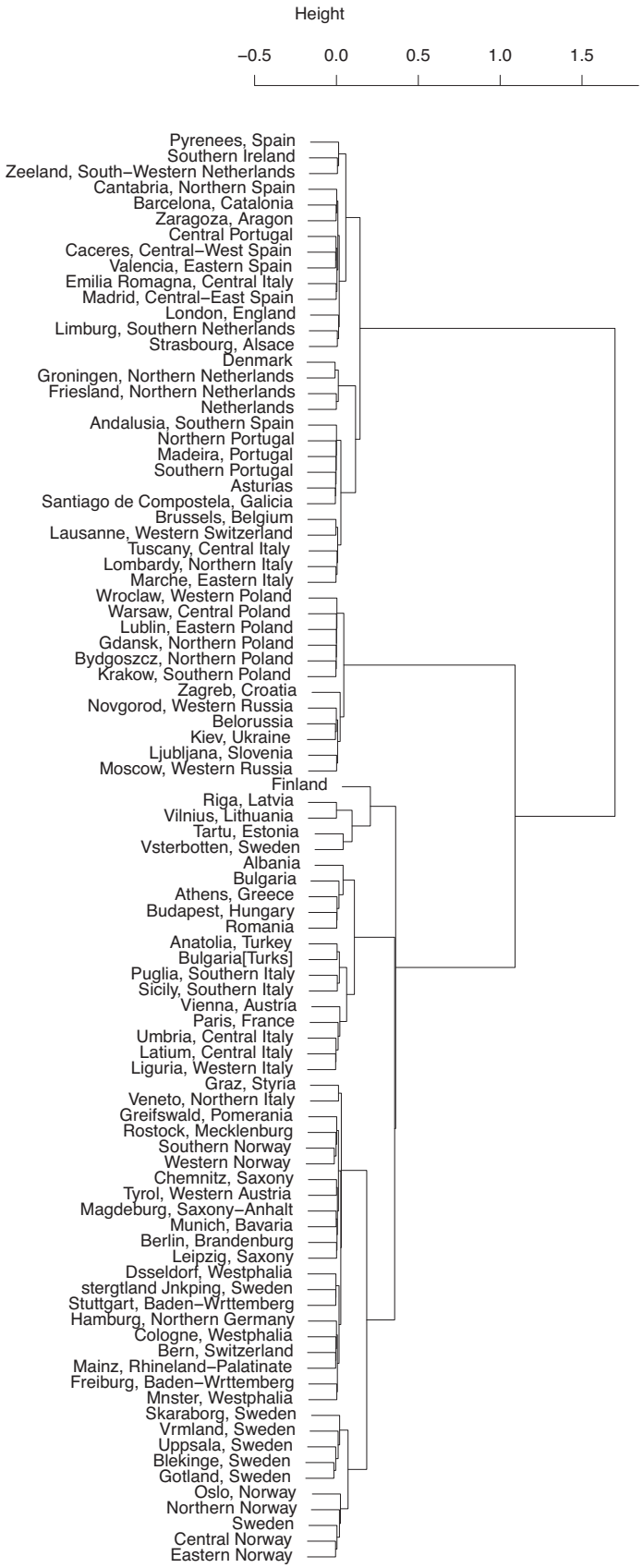

Fig. 6. A dendogram of Φ_{ST} distances [2] between all pairs of strata calculated using Arlequin version 3.5 [18]. The dendogram was made by hierarchical clustering using Ward's method [17].

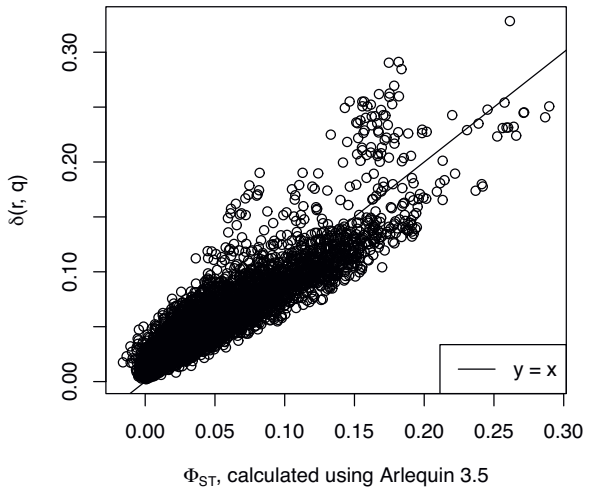

Fig. 7. Comparison of Φ_{ST} distances (from the AMOVA method of Excoffier et al. [2] calculated using Arlequin version 3.5 [18]) and the $\delta(s,t)$ distances (calculated using the discrete Laplace method). Pearson's correlation coefficient: 0.90; $p < 10^{-15}$.

$$H_s = \sum_{j=1}^{c} w_{sj} \log w_{sj} \tag{8}$$

be the homogeneity entropy of the s'th stratum for $s \in \{1, 2, \ldots, S\}$. Let

$$e_i = \sum_{j=1}^{c} \hat{v}_{ij} \log \hat{v}_{ij} \tag{9}$$

be the entropy of the i'th individual for $i \in \{1, 2, \ldots, n\}$, and let

$$P_s = n_s^{-1} \sum_{i \in I_s} e_i \tag{10}$$

be the subpopulation certainty entropy of the sth stratum for $s \in \{1, 2, \ldots, S\}$. Note, that H_s is the entropy of the $\hat{v}_{ij}$ means, whereas P_s is the mean of the $\hat{v}_{ij}$ entropies.

Three extreme strata are now chosen for further investigations. These three strata are 'Finland' (lowest homogeneity entropy $H_s = 2.29$ and lowest subpopulation certainty entropy $P_s = 0.60$), 'Lublin, Eastern Poland' (homogeneity entropy $H_s = 3.07$ and highest subpopulation certainty entropy $P_s = 1.16$) and 'Berlin, Brandenburg, Germany' (highest homogeneity entropy $H_s = 3.43$ and subpopulation certainty entropy $P_s = 0.86$).

In Fig. 8, $\hat{v}_{ij}$ values are plottet for 'Finland', 'Lublin, Eastern Poland' and 'Berlin, Brandenburg, Germany'. This can also be done for the four mega clusters (Fig. 9).

3.3. Marginals

To validate a model of the Y-STR haplotype probability distribution, a reasonable validation criterium is that the predicted single and pairwise marginal allele distributions fit well with the observed distributions. This means that if e.g. 50% of the individuals in the database have allele 14 at `DYS19` (disregarding the alleles at the other loci), then this should also be predicted by the discrete Laplace method.

3.3.1. Single marginals

For each locus, the observed marginal distribution (percentage of individuals having each allele) can be compared with the expected distribution under the discrete Laplace method that is given by

$$P(x) = \sum_{j=1}^{c} \hat{\tau}_j f(|x - \hat{y}_{jk}|; \hat{p}_{jk}) \tag{11}$$

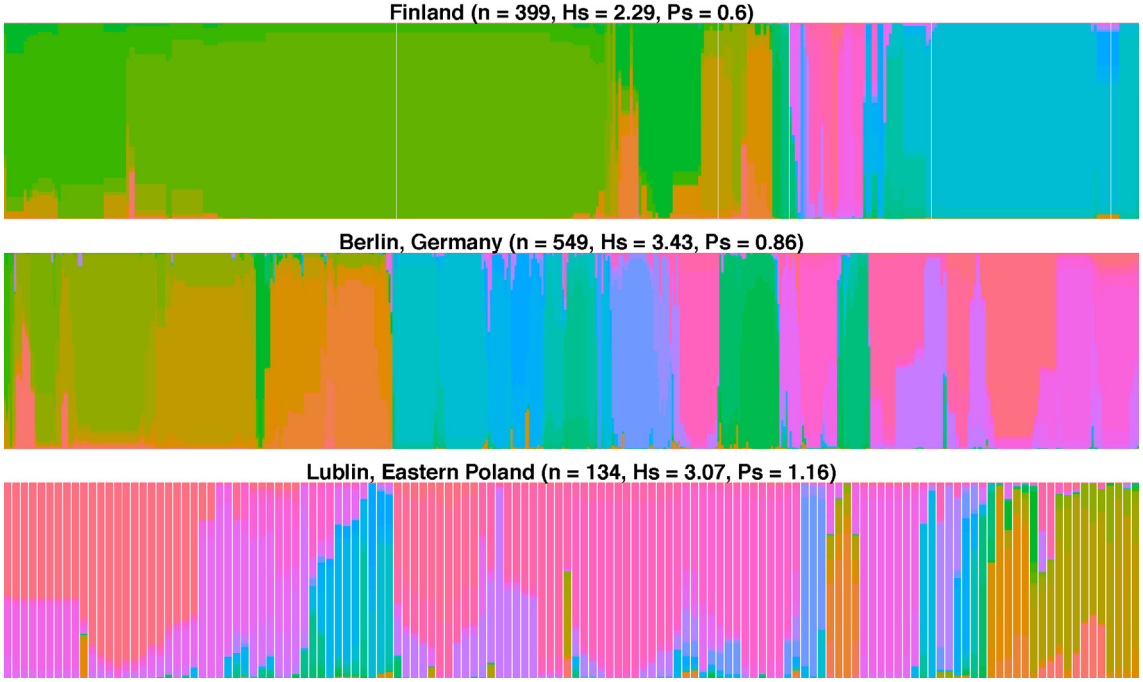

Fig. 8. Each vertical bar shows an individual's row in the v_{ij} matrix (such that the ith vertical bar consists of the 40 numbers $\{v_{ij}\}_j$ for $j \in \{1, 2, \ldots, 40\}$). The v_{ij} matrices are shown for Finland (lowest homogeneity entropy, $H_s = 2.29$, and lowest subpopulation certainty entropy, $P_s = 0.60$), Lublin, Eastern Poland (homogeneity entropy, $H_s = 3.07$, and highest subpopulation certainty entropy, $P_s = 1.16$) and Berlin, Germany (highest homogeneity entropy, $H_s = 3.43$, and subpopulation certainty entropy, $P_s = 0.86$). The subpopulations (the columns of the v_{ij} matrices) have the same order and colour as in Fig. 2. In Fig. 9, a similar figure is shown for four mega clusters. The individuals were reordered using the `R` library `seriation` [12,13] with the `BEA_TSP` method [12]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

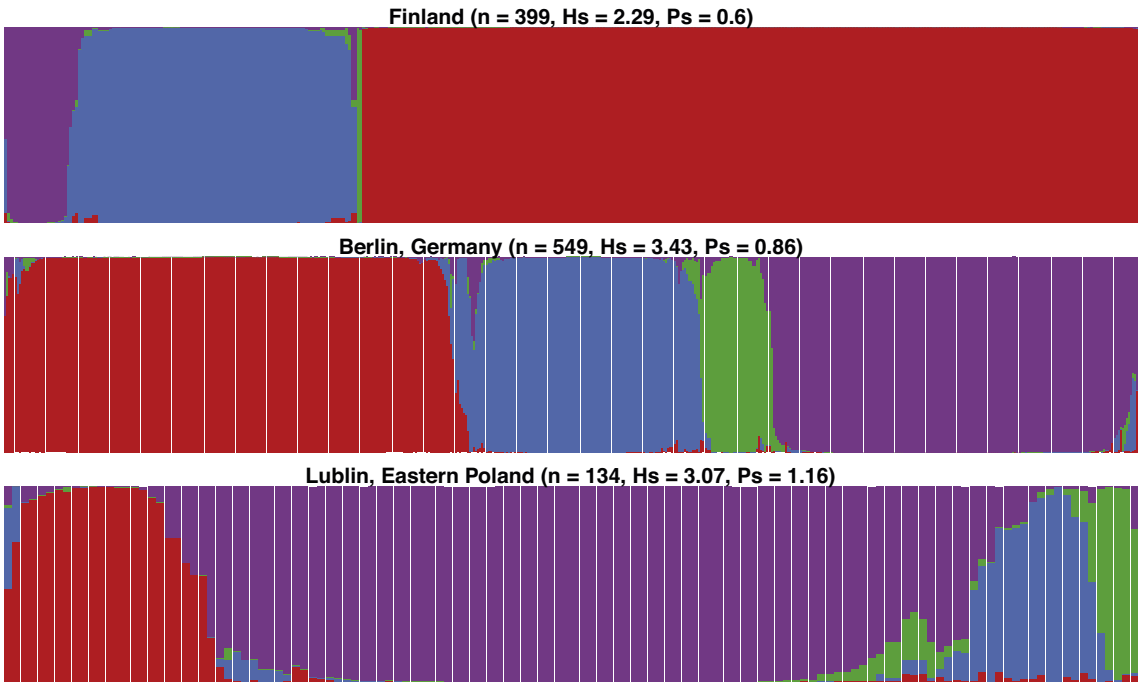

Fig. 9. Please refer to the caption of Fig. 8. The subpopulations (the columns of the merged v_{ij} matrices were reordered such that four mega clusters were obtained) have the same order and colour as in Fig. 4. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

for each allele x at the k'th locus. Fig. 10 shows the single marginal distribution for each locus. Note, that this is a mixture of discrete Laplace distributions, which means that it is not necessarily shaped like a single, discrete Laplace distribution.

3.3.2. Pairwise marginals

For two loci, k and l, the observed marginal distribution (number of individuals having each combination of alleles at the two loci) can be compared with the expected distribution under the discrete Laplace method that is given by

$$P(x_k, x_l) = \sum_{j=1}^{c} \hat{\tau}_j \, f\left(|x_k - \hat{y}_{jk}|; \hat{p}_{jk}\right) f\left(|x_l - \hat{y}_{jl}|; \hat{p}_{jl}\right) \quad (12)$$

for alleles (x_k, x_l) for locus k and l, respectively.

4. Discussion

We have demonstrated that the discrete Laplace method (analysing a mixture of multivariate, marginally independent, discrete Laplace distributions) as described by Andersen et al.

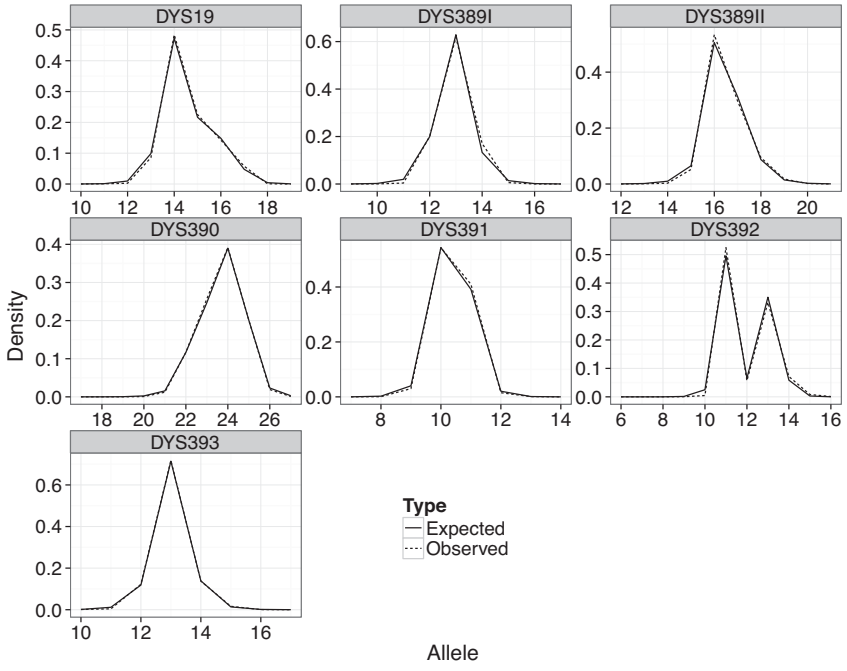

Fig. 10. Single marginal observed and expected (by the discrete Laplace method) distributions for each Y-STR locus.

[3,4] is a valuable tool for modelling Y-chromosomal STR haplotypes and for making inference based on such a modelling. The discrete Laplace method can be used for a wide range of tasks such as haplotype frequency estimation and model based cluster analysis (e.g. in analysing population structure). Furthermore, the calculations can be performed on a normal computer.

In the model based cluster analysis, Western and Eastern European subpopulations were identified (refer to Fig. 3) similar to the results of Roewer et al. [1] obtained using the AMOVA method by Excoffier et al. [2]. A more detailed map of Europe using all identified subpopulations is shown in Fig. 2.

Another comparison of the discrete Laplace method with the AMOVA method [2] was performed in Section 3.1. Here, it was shown that there was good agreement between the pairwise distances between strata (sample locations) obtained using the discrete Laplace method and the AMOVA method for both European and African population samples.

Homogeneity was analysed in two different ways, see Section 3.2. We found that the Y-STR haplotypes from Finland were more homogeneous than those from Lublin, Eastern Poland and Berlin, Germany (refer to Fig. 8). Lublin is known to have been a center for trade [19], so this heterogeneity seems quite reasonable.

In this paper, we have focused on cluster analysis on stratum level. Analysis on individual level rather than on stratum level is possible, refer to Section 2.1. Other methods for analysing on the individual level include e.g. the STRUCTURE software [23] and the ADMIXTURE software [24]. As the focus of this paper was on stratum level cluster analysis, we have not compared our results to corresponding obtained by STRUCTURE nor ADMIXTURE.

The discrete Laplace method makes it possible to calculate the expected distribution of STR alleles (expected percentage of individuals having a certain allele). We demonstrated that the expected distribution of alleles at each locus was similar to the observed distribution (refer to Section 3.3 and 10).

The discrete Laplace method was developed for and applied on haploid data. However, it may also be applicable for diploid genotypes, e.g. by having a mixture distribution for each locus or by having phased haplotypes.

Acknowledgements

Thanks to both Oticon Foundation (Grant number 13-2102) and Ellen and Aage Andersen's Foundation for supporting parts of this work.

Appendix A. Statistical model

Assume that the parameters of the discrete Laplace model [3] are estimated, e.g. by using the `R` [8] library `disclapmix` [9] that is described and demonstrated with both simple and more advanced examples in Andersen et al. [4]. The estimated parameters are:

- The number of components in the mixture that can be interpreted as the number of estimated (genetic) subpopulations, $\hat{c}$ (from now on just c for easier notation).
- The central haplotype, $\hat{y}_j = (\hat{y}_{j1}, \hat{y}_{j2}, \ldots, \hat{y}_{jr})$, of the subpopulations for $j \in \{1, 2, \ldots, c\}$. Subpopulations are constructed such that the individuals are close to the central haplotype.
- The prior probabilities, $\hat{\tau}_j$ for $j \in \{1, 2, \ldots, c\}$, of belonging to the jth subpopulation.
- The parameters of the multivariate, marginally independent, discrete Laplace distributions, $\hat{p}_{jk} = \exp(\hat{\omega}_j + \hat{\lambda}_k)$ for $j \in \{1, 2, \ldots, c\}$ and $k \in \{1, 2, \ldots, r\}$. This means that there is an additive effect of locus, $\hat{\lambda}_k$, and an additive effect of subpopulation, $\hat{\omega}_j$, as described by Andersen et al. [3].

The subpopulation membership can be formulated as

$$v_{ij} = \begin{cases} 1 & \text{if the } i\text{th individual originates from the } j\text{th subpopulation,} \\ 0 & \text{otherwise.} \end{cases} \tag{A.1}$$

Thus, an individual can originate from only one subpopulation. Because the membership is not observed, the probability of each outcome is instead estimated using the EM algorithm by Dempster et al. [20] as described by Andersen et al. [3]. Thus, given x_i, let $\hat{v}_{ij}$ be the estimated probability that the ith individual originates from the jth subpopulation. Thus, $\hat{v}_{i+} = \sum_{j=1}^{c} \hat{v}_{ij} = 1$ for all i. The estimation procedure described by Andersen et al. [3] results in

$$\hat{v}_{+j} = \sum_{i=1}^{n} \hat{v}_{ij} = \hat{\tau}_j \tag{A.2}$$

being the prior probability of belonging to the jth subpopulation.

The estimate of the parameter, c, the number of subpopulations, can be obtained by using e.g. the Bayesian information criteria (BIC) [21] for various choices of c.

For the European dataset introduced in the main text, 40 subpopulations were found to be optimal among the subpopulation counts that we used (which were from 5 to 60 at intervals of 5). This was done using the `disclapmix` library version 1.2 for the statistical software `R` [8] as shown below:

```
> library(disclapmix)
> str(db) # Note, the db must be an integer matrix
 int [1:12727, 1:7] 12 12 13 13 13 13 13 13 13 13 ...
 - attr(*, "dimnames")=List of 2
  ..$ : NULL
  ..$ : chr [1:7] "DYS19" "DYS389I" "DYS389II" "DYS390" ...
> head(db)
     DYS19 DYS389I DYS389II DYS390 DYS391 DYS392 DYS393
[1,]    12      13       17     24     10     11     13
[2,]    12      13       17     24     10     11     14
[3,]    13      12       18     24     10     11     13
[4,]    13      13       16     23     10     11     13
[5,]    13      13       16     24     10     11     14
[6,]    13      13       16     24     11     13     13
> head(popnames)
[1] Albania Albania Albania Albania Albania Albania
91 Levels: Albania Anatolia,Turkey Andalusia,Southern_Spain ... Zeeland,South-
    Western_Netherlands
> fits <- lapply(seq(5L, 60L, 5), function(clusters) disclapmix(db, clusters = clusters))
> fits_BIC <- sapply(fits, function(fit) fit$BIC_marginal)
> bestfit <- fits[[which.min(fits_BIC)]]
> summary(bestfit)
```

Please, see the values of the marginal BICs in Table A.1.

Table A.1
The values of the marginal BICs at subpopulation counts from 5 to 60 at intervals of 5.

Subpopulations	BIC
5	196,525
10	187,973
15	183,595
20	182,216
25	181,408
30	180,646
35	180,532
40	180,525
45	180,583
50	180,556
55	180,552
60	180,735

A.1. Haplotype frequency estimation

Haplotype frequency estimation using the discrete Laplace method was performed as follows: Given the central haplotype of the subpopulations, $\hat{y}_j$ for $j \in \{1, 2, \ldots, c\}$, parameters $\hat{p}_{jk}$ for $j \in \{1, 2, \ldots, c\}$ and $k \in \{1, 2, \ldots, r\}$ and prior probabilities $\hat{\tau}_j$ for $j \in \{1, 2, \ldots, c\}$, the haplotype frequency of a haplotype $h = (h_1, h_2, \ldots, h_r)$ with $h_k \in \mathbb{Z}$ for $k \in \{1, 2, \ldots, r\}$ was estimated as

$$\sum_{j=1}^{c} \hat{\tau}_j \prod_{k=1}^{r} f(|h_k - \hat{y}_{jk}|; \hat{p}_{jk}), \tag{A.3}$$

where

$$f(d; p) = \left(\frac{1-p}{1+p}\right) p^{|d|} \tag{A.4}$$

is the probability mass function of a discrete Laplace distribution with parameter $0 < p < 1$ evaluated at $d \in \mathbb{Z}$.

This can be done using the `disclapmix` library for all haplotypes in the dataset:

```
estimates <- predict(bestfit, newdata = db)
```

A.2. Model based clustering

Refer to the notation of Section 2.1 of the main text. The w_{sj} were calculated in R as shown below:

```
> vij <- bestfit$v_matrix
> wsj <- aggregate(vij, list(popnames), mean)
> rownames(wsj) <- wsj[, 1]
> wsj <- as.matrix(wsj[, -1])
```

The majority of the central haplotypes of the subpopulations were close to each other. To better visualise the subpopulations, those with central haplotypes close to each other can be assembled into mega clusters as follows. Given a desired number of mega

clusters, T, let J_t for $t \in \{1, 2, \ldots, T\}$ be a partition of $\{1, 2, \ldots, c\}$ such that

$$\bigcup_{t=1}^{T} J_t = \{1, 2, \ldots, c\} \quad \text{and} \quad \bigcap_{t=1}^{T} J_t = \emptyset, \tag{A.5}$$

where $\emptyset = \{\}$ is the empty set. The collapsed w_{sj} values are

$$u_{st} = \sum_{j \in J_t} w_{sj}, \tag{A.6}$$

such that

$$u_{+t} = \sum_{s=1}^{S} u_{st} = \sum_{s=1}^{S} \sum_{j \in J_t} w_{sj} = \sum_{j \in J_t} \sum_{s=1}^{S} w_{sj} = \sum_{j \in J_t} \hat{\tau}_j \tag{A.7}$$

for the tth cluster and

$$u_{s+} = \sum_{t=1}^{T} u_{st} = \sum_{t=1}^{T} \sum_{j \in J_t} w_{sj} = 1 \tag{A.8}$$

for the sth stratum.

This means that we add together subpopulations J_t by adding their respective w_{sj} values for $j \in J_t$ to obtain mega clusters. Note, that the strata (or information about strata) are not used for constructing the mega clusters; only the central haplotype of the subpopulations and $\hat{p}_{jk}$ parameters are used.

Appendix B. Kullback–Leibler distance measure

Let $f(d\,;\,p)$ be the probability mass function of the discrete Laplace distribution. For the kth locus together with subpopulation g and h, let

$$z_1 = \hat{y}_{gk}, \quad z_2 = \hat{y}_{hk}, \quad p_1 = \hat{p}_{gk} \quad \text{and} \quad p_2 = \hat{p}_{hk}, \tag{B.1}$$

such that the distance from subpopulation g to h can be defined as

$$\mathrm{KL}'_k(g,h) = \sum_{d \in \mathbb{Z}} f(|d - z_1|;\, p_1) \log\left(\frac{f(|d - z_1|;\, p_1)}{f(|d - z_2|;\, p_2)}\right) \tag{B.2}$$

$$= \sum_{d \in \mathbb{Z}} \left(\frac{1 - p_1}{1 + p_1}\right) p_1^{|d - z_1|} \log\left(\frac{\left(\frac{1-p_1}{1+p_1}\right) p_1^{|d-z_1|}}{\left(\frac{1-p_2}{1+p_2}\right) p_2^{|d-z_2|}}\right) \tag{B.3}$$

$$= \left(\frac{1 - p_1}{1 + p_1}\right) \tag{B.4}$$

$$\times \sum_{d \in \mathbb{Z}} p_1^{|d - z_1|} \left\{ |d - z_1| \log p_1 + \log\left(\frac{1 - p_1}{1 + p_1}\right) - |d - z_2| \log p_2 - \log\left(\frac{1 - p_2}{1 + p_2}\right) \right\} \tag{B.5}$$

$$= \left(\frac{1 - p_1}{1 + p_1}\right) \left(\mathrm{KL}_k^{(1)}(g,h) + \mathrm{KL}_k^{(2)}(g,h) + \mathrm{KL}_k^{(3)}(g,h) + \mathrm{KL}_k^{(4)}(g,h) \right), \tag{B.6}$$

where

$$\mathrm{KL}_k^{(1)}(g,h) = \sum_{d \in \mathbb{Z}} p_1^{|d - z_1|} |d - z_1| \log p_1 = \log p_1 \sum_{d \in \mathbb{Z}} |d|\, p_1^{|d|} = 2 \log p_1 \sum_{d=1}^{\infty} d\, p_1^d = \frac{2\, p_1 \log p_1}{(p_1 - 1)^2} \tag{B.7}$$

$$\mathrm{KL}_k^{(2)}(g,h) = \sum_{d \in \mathbb{Z}} p_1^{|d - z_1|} \log\left(\frac{1 - p_1}{1 + p_1}\right) = \log\left(\frac{1 - p_1}{1 + p_1}\right) \sum_{d \in \mathbb{Z}} p_1^{|d|} = \log\left(\frac{1 - p_1}{1 + p_1}\right) \left(1 + 2 \sum_{d=1}^{\infty} p_1^d\right) \tag{B.8}$$

$$= \log\left(\frac{1 - p_1}{1 + p_1}\right) \left(1 + \frac{2\, p_1}{1 - p_1}\right) = \left(\frac{1 + p_1}{1 - p_1}\right) \log\left(\frac{1 - p_1}{1 + p_1}\right) \tag{B.9}$$

$$\mathrm{KL}_k^{(3)}(g,h) = -\sum_{d \in \mathbb{Z}} p_1^{|d - z_1|} |d - z_2| \log p_2 = -\log p_2 \sum_{d \in \mathbb{Z}} |d - z_2|\, p_1^{|d - z_1|} \tag{B.10}$$

$$= -\log p_2 \sum_{d \in \mathbb{Z}} |d - z_2 + z_1|\, p_1^{|d|} = -\log p_2 \sum_{d \in \mathbb{Z}} |z_2 - z_1 - d|\, p_1^{|d|} \tag{B.11}$$

$$\mathrm{KL}_k^{(4)}(g,h) = -\sum_{d \in \mathbb{Z}} p_1^{|d - z_1|} \log\left(\frac{1 - p_2}{1 + p_2}\right) = -\log\left(\frac{1 - p_2}{1 + p_2}\right) \sum_{d \in \mathbb{Z}} p_1^{|d|} = -\left(\frac{1 + p_1}{1 - p_1}\right) \log\left(\frac{1 - p_2}{1 + p_2}\right). \tag{B.12}$$

To evaluate $\mathrm{KL}_k^{(3)}(g,h)$, note that

$$\sum_{d\in\mathbb{Z}} |z_2 - z_1 - d|\, p_1^{|d|} = \sum_{d\in\mathbb{Z}} |-(z_2 - z_1) + d|\, p_1^{|d|} \tag{B.13}$$

$$= \sum_{-d\in\mathbb{Z}} |-(z_2 - z_1) - d|\, p_1^{|d|} \tag{B.14}$$

$$= \sum_{-d\in\mathbb{Z}} |z_2 - z_1 + d|\, p_1^{|d|}, \tag{B.15}$$

such that for $m = |z_2 - z_1| \geq 0$,

$$\sum_{d\in\mathbb{Z}} |z_2 - z_1 - d|\, p_1^{|d|} = \sum_{d\in\mathbb{Z}} |m - d|\, p_1^{|d|} \tag{B.16}$$

$$= \sum_{d=-\infty}^{0} |m - d|\, p_1^{|d|} + \sum_{d=1}^{m} |m - d|\, p_1^{|d|} + \sum_{d=m+1}^{\infty} |m - d|\, p_1^{|d|} \tag{B.17}$$

$$= \sum_{d=0}^{\infty} (m + d)\, p_1^d + \sum_{d=1}^{m} (m - d)\, p_1^d + \sum_{d=m+1}^{\infty} (d - m)\, p_1^d \tag{B.18}$$

$$= \frac{(1 - p_1)m + p_1}{(p_1 - 1)^2} + \frac{p_1(p_1^m - m(p_1 - 1) - 1)}{(p_1 - 1)^2} + \frac{p^{m+1}}{(p_1 - 1)^2} \tag{B.19}$$

$$= \frac{2\,p_1^{m+1} - m(p_1^2 - 1)}{(p_1 - 1)^2}, \tag{B.20}$$

resulting in

$$\mathrm{KL}_k^{(3)}(g,h) = -\left(\frac{2\,p_1^{m+1} - m(p_1^2 - 1)}{(p_1 - 1)^2}\right)\log p_2. \tag{B.21}$$

Thus,

$$\mathrm{KL}'_k(g,h) = \left(\frac{1 - p_1}{1 + p_1}\right)\frac{2\,p_1 \log p_1}{(1 - p_1)^2} + \log\left(\frac{1 - p_1}{1 + p_1}\right) \tag{B.22}$$

$$- \left(\frac{1 - p_1}{1 + p_1}\right)\left(\frac{2\,p_1^{m+1} - m(p_1^2 - 1)}{(p_1 - 1)^2}\right)\log p_2 - \log\left(\frac{1 - p_2}{1 + p_2}\right) \tag{B.23}$$

$$= \frac{2\,p_1 \log p_1}{1 - p_1^2} + \log\left(\frac{(1 - p_1)(1 + p_2)}{(1 + p_1)(1 - p_2)}\right) - \left(\frac{2\,p_1^{m+1} - m(p_1^2 - 1)}{1 - p_1^2}\right)\log p_2. \tag{B.24}$$

To make the distance symmetric, let

$$\mathrm{KL}_k(g,h) = \mathrm{KL}'_k(g,h) + \mathrm{KL}'_k(h,g). \tag{B.25}$$

Because mutations are assumed to happen independently across loci, we can sum the distances at each locus such that

$$\mathrm{KL}(g,h) = \sum_{k=1}^{r} \mathrm{KL}_k(g,h) \tag{B.26}$$

is the distance between subpopulation g and h.

References

[1] L. Roewer, P.J.P. Croucher, S. Willuweit, T.T. Lu, M. Kayser, R. Lessig, P. de Knijff, M.A. Jobling, C. Tyler-Smith, M. Krawczak, Signature of recent historical events in the European Y-chromosomal STR haplotype distribution, Hum. Genet. 116 (2005) 279–291.

[2] L. Excoffier, P.E. Smouse, J.M. Quattro, Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data, Genetics 131 (1992) 479–491.

[3] M.M. Andersen, P.S. Eriksen, N. Morling, The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies, J. Theor. Biol. 329 (2013) 39–51.

[4] M.M. Andersen, P.S. Eriksen, N. Morling, A gentle introduction to the discrete Laplace method for estimating Y-STR haplotype frequencies, 2013 arXiv:1304.2129.

[5] J. Buckleton, M. Krawczak, B. Weir, The interpretation of lineage markers in forensic DNA testing, Forensic Sci. Int. Genet. 5 (2011) 78–83.

[6] M.M. Andersen, A. Caliebe, A. Jochens, S. Willuweit, M. Krawczak, Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory, Forensic Sci. Int. Genet. 7 (2013) 264–271.

[7] C.H. Brenner, Fundamental problem of forensic mathematics – the evidential value of a rare haplotype, Forensic Sci. Int. Genet. 4 (2010) 281–291.

[8] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2012, ISBN 3-900051-07-0.

[9] M.M. Andersen, P.S. Eriksen, disclapmix: Discrete Laplace Mixture Inference using the EM Algorithm, 2013 R package version 1.2.

[10] T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, Biol. Skrifter 5 (1948) 1–34.

[11] D. Defays, An efficient algorithm for a complete link method, Comput. J. 4 (1977) 364–366.
[12] M. Hahsler, K. Hornik, C. Buchta, Getting things in order: an introduction to the R package seriation, J. Stat. Softw. 25 (2008) 1–34.
[13] M. Hahsler, C. Buchta, K. Hornik, Infrastructure for Seriation, 2012 R package version 1.0-7.
[14] L. Roewer, M. Krawczak, S. Willuweit, et al., Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes, Forensic Sci. Int. 2-3 (2001) 106–113.
[15] S. Willuweit, L. Roewer, Y chromosome haplotype reference database (YHRD): update, Forensic Sci. Int. Genet. 1 (2009) 83–87.
[16] J.M. Butler, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, 2nd ed., Academic Press, 2005.
[17] J.H. Ward Jr., Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 48 (1963) 236–244.
[18] L. Excoffier, H.L. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (2010) 564–567.
[19] G.J. Lerski, Historical Dictionary of Poland, 966–1945, Greenwood, Westport, CT, USA, 1996.
[20] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. 39 (1977) 1–38.
[21] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[22] C. de Filippo, C. Barbieri, M. Whitten, S.W. Mpoloka, E.D. Gunnarsdóttir, K. Bostoen, T. Nyambe, K. Beyer, H. Schreiber, P. de Knijff, D. Luiselli, M. Stoneking, B. Pakendorf, Y-chromosomal variation in Sub-Saharan Africa: insights into the history of Niger-Congo groups, Molecular Biology and Evolution (2010).
[23] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2) (2000) 945–959.
[24] D.H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals, Genome Research 19 (2009) 1655–1664.